Rethinking Layered Graphic Design Generation with a Top-Down Approach – Appendix –

A. Gallery for more text-to-template results

We demonstrate more text-to-template results in Figure 1 and Figure 2. Due to the text being layered and vectorized, operations such as selection and copying are feasible.

B. Prompts for the stage1 reference creation

Below we use in-context learning to make VLM expand the given prompt:

> Your task is to expand the original prompt into a detailed one. I will give you some examples.
> [Input 1] Create an advertisement for a fish market with a special offer of a 20% discount on seafood.
> [Output 1] A bustling fish market under a vibrant morning sky. Local vendors display an array of fresh, glistening seafood, from ruby-red lobsters to iridescent, silver fish. A large, colorful banner hangs overhead, proudly announcing a special offer with 20% discount on all seafood. The air is thick with excitement and the irresistible aroma of the ocean.
> [Input 2] Create a business card for a flower shop with a focus on blue tulips.
> [Output 2] An elegant business card lying on a white marble surface. The card is adorned with a captivating watercolor illustration of rich, azure blue tulips, their petals opening up to reveal layers of deep and light shades of blue. The shop's name is written in a sophisticated cursive font at the center, while contact details are subtly placed at the bottom right corner.
> Now based on the given prompt "Design a cutting-edge logo for a real estate agency named Golden Home.", please expand it into a detailed one.

When users wish to provide additional constraints, they can input sketches to the VLM. We employ the following task prompt template to query the VLM:

> You will be provided with a sketch that you need to analyze and describe meticulously, paying close attention to each detail depicted. Identify and describe where each object is located within the sketch. Note that the "xxx" symbols on the image are placeholders for text, which you should replace with appropriate content. Your description should capture the layout and the thematic elements of the design. As a reference information, this image is about "eating more apples is good for your health".

This template explicitly describes the task and indicates that specific text should replace the placeholder 'xxx'. It also includes a brief user intention, providing the VLM with an overview of the image. We demonstrate some cases in Figure 3.

C. Combined prompts for the stage2 design planning

We showcase some examples in the below. For training GenAI design, we use the following template:

> Parse and refine the attributions of text. Parse the objects, and backgrounds in the graphic design image. The caption of the image is The "Red White Bold Type" beverage label is a striking visual feast, designed to capture the essence of boldness and purity. With a vivid red and pristine white color scheme, the label features bold, assertive typography that commands attention. This design not only reflects the vibrant and robust flavors of the beverage but also appeals to

Figure 1. Visualizations of more text-to-template results. [1/2]

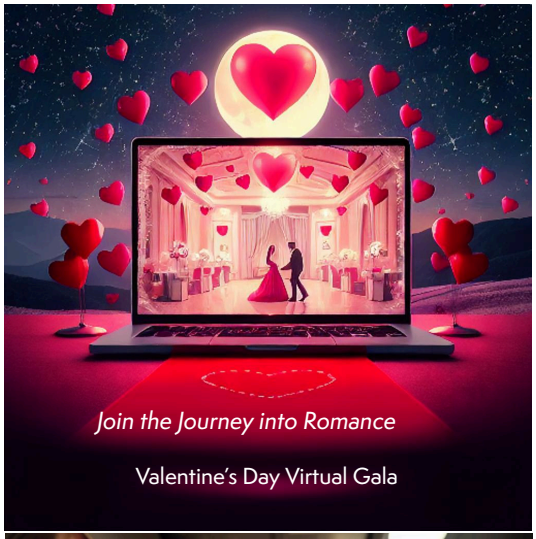


Figure 2. Visualizations of more text-to-template results. [2/2]

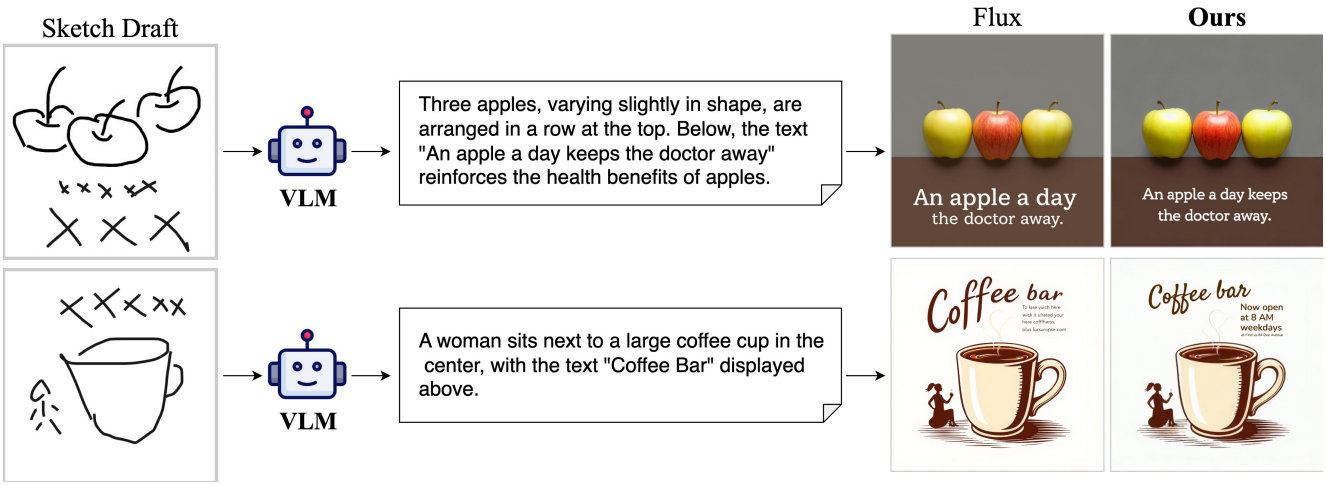


Figure 3. The VLM can convert the sketch draft into detailed prompt to generate references.

consumers with its clean, contemporary aesthetic, making it a standout choice on any shelf. Support OCR results are: [[(22, 64, 228, 132)], [(21, 126, 311, 211)], [(82, 208, 119, 215)]].

Parse and refine the attributions of text. Parse the objects, and backgrounds in the graphic design image. The caption of the image is The Facebook page cover for a modern record store should be a vibrant and engaging visual that encapsulates the essence of music and contemporary design. It might feature a collage of iconic album covers, interspersed with sleek, modern graphic elements that convey the store's cutting-edge aesthetic. Support OCR results are: [[(214, 89, 299, 120)], [(41, 86, 110, 138)], [(18, 121, 59, 176)], [(195, 121, 317, 147)], [(209, 175, 310, 197)], [(224, 197, 290, 219)], [(84, 219, 106, 237)], [(215, 232, 300, 246)]].

For training the original design, we use the following template. Please note that here we do not incorporate the description since the text within the design already contains massive information. Meanwhile, we integrate the OCR recognition result in the OCR string.

Parse the attributions of text, objects, and backgrounds in the graphic design image. Support OCR results are: [['THE COOD', (85, 15, 228, 51)], ['CREATIVE', (88, 51, 232, 85)], ['STUDIO', (84, 83, 196, 120)], ['2701Willow', (85, 218, 158, 236)], ['Charles,', (83, 235, 135, 253)], ['aneLake', (122, 228, 177, 243)], ['(555)555-0100', (85, 265, 174, 282)], ['@thegoodstudio', (86, 297, 180, 312)], ['www.thegoodstudio.site.con', (85, 310, 237, 324)]]

Parse the attributions of text, objects, and backgrounds in the graphic design image. Support OCR results are: [['CLEARANCE', (19, 213, 318, 255)], ['SALE', (14, 256, 136, 297)], ['2701WillowOaks', (203, 272, 300, 287)], ['Lane Lake Charles,LA', (203, 284, 321, 298)]]

For training the backgrounds with text, we use the following template. Note that the OCR string is omitted since there is no text within the design.

Add text on the background. And parse the overall graphic design. The caption of the image is floral green and pink wellness institute business card.

Add text on the background. And parse the overall graphic design. The caption of the image is The logo for Green Saw Carpenters captures the essence of the brands commitment to sustainable building practices and skilled craftsmanship. It features a stylized green saw blade, intricately designed to resemble both a leaf and a carpentry tool, symbolizing the fusion of nature and construction.

D. More details about the questionnaire for result selection

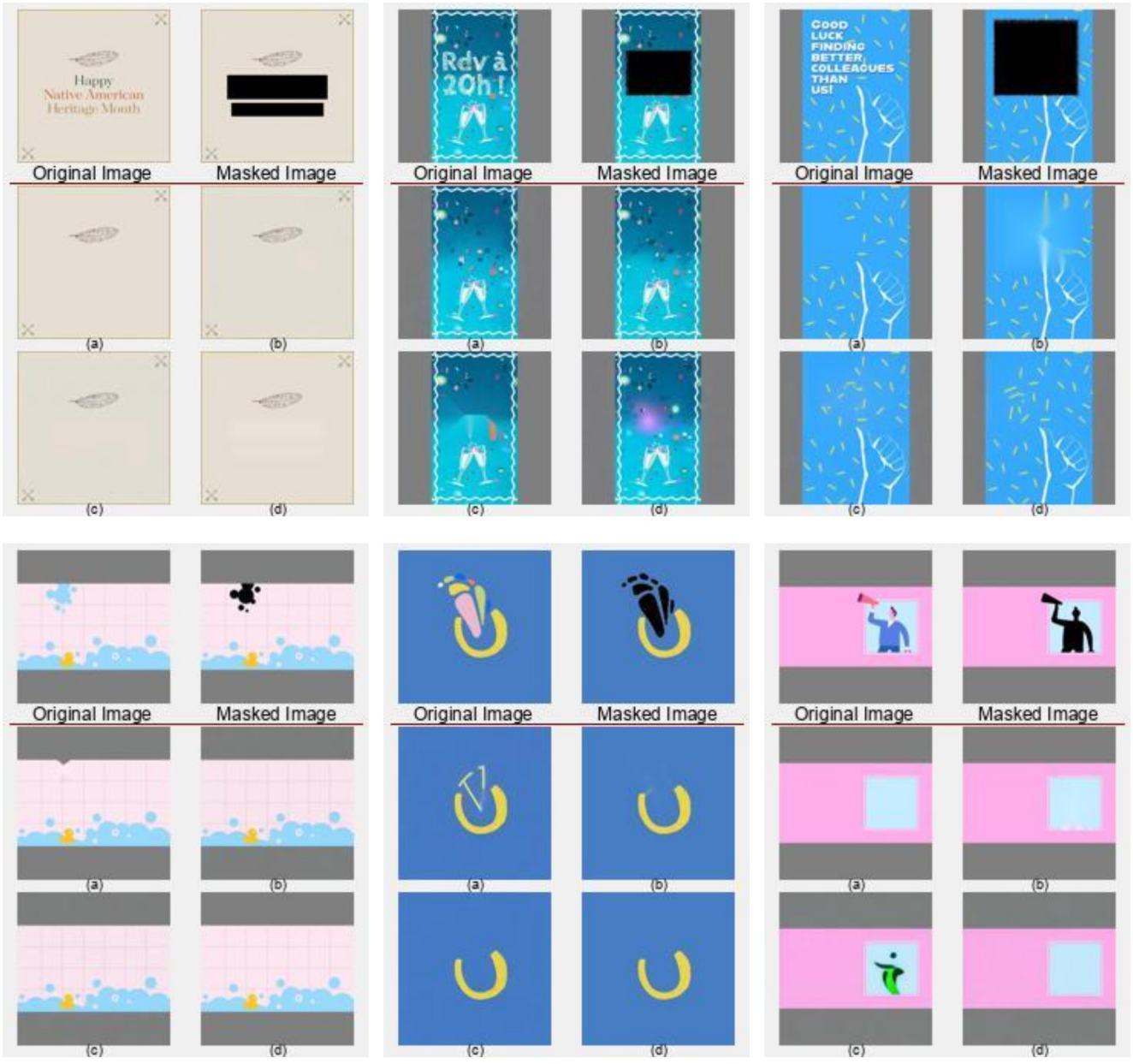


Figure 4. More examples about the questionnaire. The samples in the first row is for text removal, and the samples in the second row is for object removal.

We show the task prompt in the following, and display some cases in Figure 4.

The provided image appears to show four different results of a graphic design removal task. The first row displays the original image on the left and the masked image on the right. The second and third rows exhibit the corresponding outcomes of the graphic design removal. To evaluate the effectiveness of the results, the key criteria are: 1) the overall harmony and coherence of the image, 2) the purity and cleanness of the background, and 3) the absence of any additional, extraneous elements. Based on these criteria, please select the option (a, b, c, or d) that represents the best result.

E. More details about Design39K

Here we provide additional statistics about the in-house Design39K dataset. The average output sequence length is 728.32. To ensure that the majority of cases do not exceed this limit, we have set the maximum output length for the VLM to 1,536. Each design, on average, contains about 1.02 objects and 3.11 text regions. Notably, the dataset utilizes the title of each design as the description.

F. More details about each vision expert

Flux. We use the open-source Flux.1 schnell to create the reference. We observe that, even without explicit character guidance like other methods [1, 7], Flux still synthesizes high-quality design references. Flux demonstrates robust performance in generating reference images compared with previous methods such as SD 1.5 [6]. In some cases, Flux may struggle to generate design images according to the prompts, instead producing backgrounds without text. In these situations, users may need to attempt multiple times to obtain the desired reference image. The output size of the model is 1024×1024. The number of sampling steps is set to 4, and the maximum length of the input prompt is set to 256, both as default.
PaddleOCR. PaddleOCR is an open-source optical character recognition toolkit that provides practical and efficient solutions for text detection and recognition across various images. It has been observed that PaddleOCR also exhibits strong detection performance on GenAI images.
SAM. Segment Anything Model [5] is an advanced segmentation model designed to perform highly accurate and versatile image segmentation across an extensive array of objects and scenes, enabling detailed and automated analysis of visual data. Here we employ the detection box to obtain the segmentation mask. Specifically, we employ the "sam-vit-base" architecture to get the segmentation mask.
Removal model. We achieve object removal results using the ControlNet inpainting model based on SD 1.5 [6], employing the prompt "nothing in the image" to erase specific content. The input size and the output size are 512×512. While we notice that a few samples exhibit color shifting, we consider this acceptable as the results still appear harmonious.

G. Details about the Streamlit frontend

We develop a Streamlit-based front-end system to facilitate the presentation and manipulation of layered graphic designs. This system enables the separation and individual rendering of various design elements, including text, images, and objects, thereby allowing for flexible control and real-time previewing of design components. Specifically, we leverage HTML and CSS to render text elements and employ the st.elements.image component to display images and objects.

H. Capability to produce non-square layered design

Please note that the proposed Accordion method can produce non-layered designs, which sets it apart from existing methods like COLE [4] and Open-COLE [3] that are confined to layered designs. For instance, when a non-square design is used as a reference in the second stage, it is initially padded with gray borders to facilitate the design planning process. Subsequently, the SAM and removal models are applied to the padded image. Finally, the borders are deleted to yield the non-square designs. The results are shown in Figure 5.

I. Task prompt for evaluating adding text to background

Here we use the following prompt for GPT-4V. The horizontal concatenation is demonstrated in Figure 6.

Here you will see two designs with the same background but different text content and placement. Please consider the

Figure 5. Visualizations of non-square designs.

factors such as layout, content relevance, typography and color scheme. Select which one is better, you can answer the left one is better / the right one is better, then detail the reasons.

Figure 6. Samples generated by COLE and ours are concatenated for comparison.

J. Details of user studies by designers

In our user studies, we present several cases to evaluate the effectiveness of converting sketches to design images (Figure 7) and text to templates 8. We specifically invite individuals with design backgrounds to participate in these studies, ensuring

that the assessments are informed by professional insights. This approach enables us to gather valuable feedback on the usability and accuracy of our models from experts in the field, enhancing the validity of our study results.

Prompt: "take care of the flower"

Description:
The sketch presents a single, elegant flower depicted in a minimalist line drawing style, positioned prominently in the upper portion of the image. It features a stem, leaves, and a clearly defined blossom. Below the flower, there are two lines of text. The top line reads "Care for Flowers," and the bottom line states "Grow Beauty." These captions effectively convey the theme of nurturing and appreciating the beauty of flowers, aligning with the message of taking care of plants.

(a) The image description is reasonable; the generated image corresponds to the sketch.

(b) The image description is reasonable; the generated image DOES NOT correspond to the sketch.

(c) The image description is NOT reasonable; the generated image corresponds to the sketch.

(b) The image description is NOT reasonable; the generated image DOES NOT correspond to the sketch.

Figure 7. User studies by designers to assess to task of sketch to design images.

K. Details about the experiment on the benefits of joint training

In Table 1, we present the evaluation results using various types of images as inputs. In Table 1, we present the evaluation results using various types of images as inputs. Note that we use normalized edit distance (NED) for evaluating OCR accuracy, which is particularly effective when our OCR is applied to paragraph-level long text following [2]. For color accuracy, we consider the predictions correct only when the values for red (R), green (G), blue (B), and alpha (A) channels are exactly accurate. When the input images are designs with nonsensical text or backgrounds without text, it becomes challenging to assess text-related metrics. Therefore, we rely solely on object-related metrics for evaluation. The results indicate that the scores for both categories are comparable. Given our goal to develop a compact model, we ultimately opt for joint training.

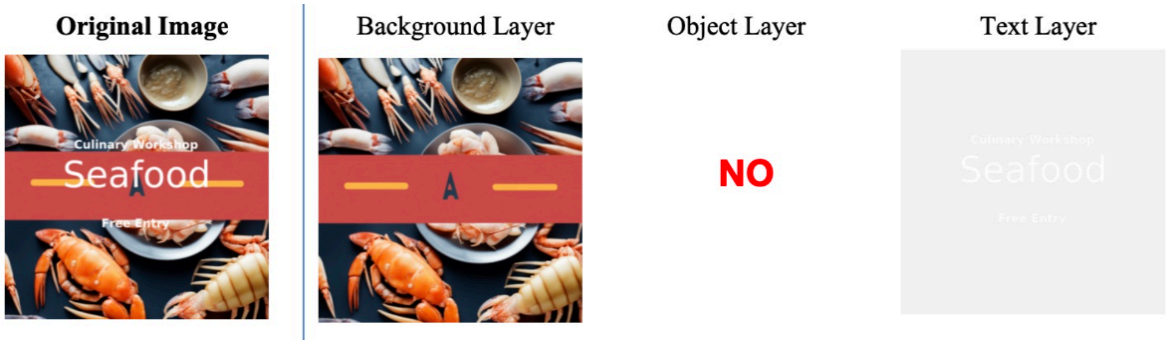
We also evaluate the effect of integrating an OCR prompt into our system. For the task of text recognition aimed at parsing the original design, we note a 7.23% improvement in paragraph-level OCR Normalized Edit Distance (NED), increasing from 61.28% to 68.51%. Additionally, the average detection F1 score for the text detection task in both the original and GenAI designs has improved by 5.46%, rising from 73.12% to 78.59%.

References

[1] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *ECCV*, 2024. 6

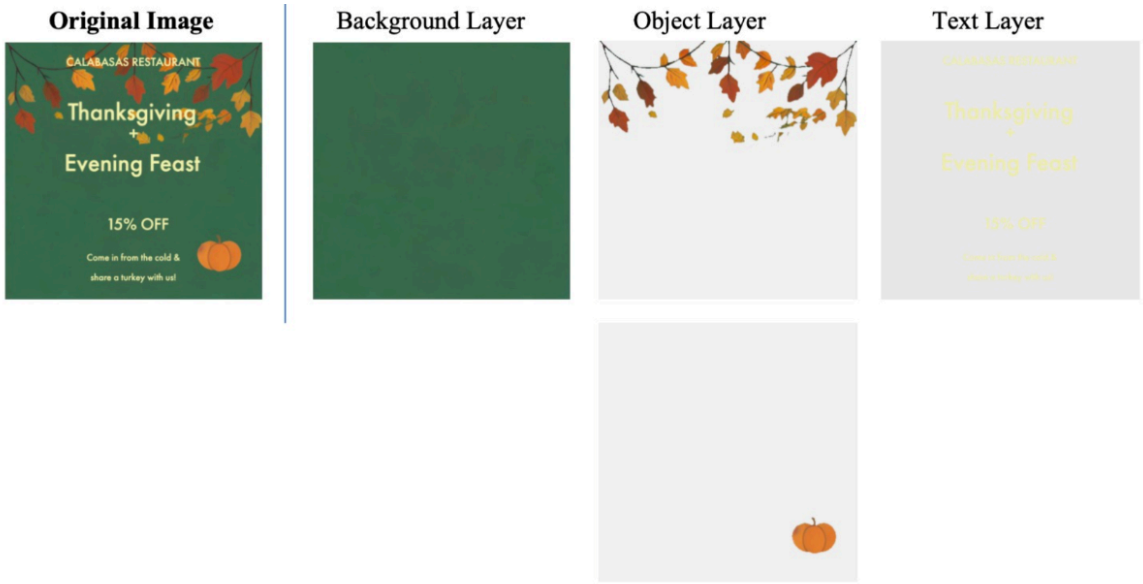
[2] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, 2019. 8

[3] Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. Opencole: Towards reproducible automatic graphic design generation. *arXiv preprint arXiv:2406.08232*, 2024. 6



(a) This layering method helps further editing.

(b) This layering method DOES NOT help further editing.



(a) This layering method helps further editing.

(b) This layering method DOES NOT help further editing.

Figure 8. User studies by designers to assess to task of text-to-templates.

[4] Peidong Jia, Chenxuan Li, Zeyu Liu, Yichao Shen, Xingru Chen, Yuhui Yuan, Yinglin Zheng, Dong Chen, Ji Li, Xiaodong Xie, et al. Cole: A hierarchical generation framework for graphic design. *arXiv preprint arXiv:2311.16974*, 2023. 6

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 6

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 6

[7] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *ICLR*, 2024. 6

Table 1. Ablation studies about the experiment on the benefits of joint training.

Metrics	Separate Training	Joint Training
Original Design		
Text Detection F1	75.42	**78.59**
Text Recognition NED	**72.87**	68.51
Object Detection F1	82.17	**84.64**
Color Accuracy	26.66	**28.09**
Font Accuracy	**24.51**	21.62
Line Number Accuracy	**86.96**	86.28
Alignment Accuracy	87.28	**88.60**
Angle Accuracy	90.16	**91.52**
Designs with Nonsensical Text		
Object Detection F1	79.27	**83.06**
Backgrounds without Text		
Object Detection F1	83.52	**86.94**
Questionnaire Result Selection		
Selection Accuracy	**83.54**	**83.54**
Average Score	72.03	**72.85**